1997 Crime Mapping Conference
Exploring the Future of Crime Mapping
Advanced Cluster Analysis

Dennis Roncek, University of Nebraska - Omaha

Numerically-based technique for classifying into groups (identify that members of a group resemble each other more than they resemble something else)

Traditional cluster analysis isn't widely used for spatial analysis research applications, although some practitioners do use it (such as Phil Canter).

Issues and concerns with current cluster mechanisms:

- Regression/hot spot clusters cannot differentiate among individuals in a multiple offender group
- Geographic profiling cannot describe multiple offenders

Traditional cluster analysis is more appropriate if **multiple groups** are believed to have committed crimes--if groups of crimes are similar to each other regardless of where they occur

Cluster analysis is the obverse of factor analysis -
Cluster analysis: Use the similarities of characteristics to identify objects
Factor analysis: Use the similarities of objects to identify characteristics

There are 7 basic families of clustering methods, 2 of which were discussed in this session

1) Hierarchical agglomerate methods - successively placing phenomena in groups based on similarity

- begin with a similarity matrix (table with events on rows and events on columns and tells similarity between the two events)
- computer searches table, finds 2 events with most similar characteristics

What kind of **similarity measure** to use?
All begin by joining the 2 most similar events, but differ in the decision rule of adding another case

Simple linkage (the simplest measure) -- "nearest neighbor"

- decision rule of adding another case: just needs to be similar to one case in the cluster
- tends to find elongated clusters (long and narrow, not circular patterns)

Complete linkage (opposite of simple linkage) -- "farthest neighbor"

- to add new cases, must be a minimum distance from every member of the group

Average linkage -- an average of existing clusters is used as the basis of similarity

Ward's method -- smallest possible increase of the error sum of squares

- minimizing the variance within the cluster
- sounds good, but very sensitive to measurement scales
- large values, like income, can have a large effect
- tends to yield circular patterns

Each method has certain general outcome tendencies
Best solution: use multiple strategies and hope similar patterns show up

2) Interactive Partitioning - Successively dividing phenomena into smaller groups based on similarity

faster than hierarchical agglomeration, good for very large data sets (recommended by SAS manuals)

Simple matching coefficient - increase score if two items have same value of an attribute
a = number of cases where both events have characteristic
b = number of cases where both events do not have characteristic
c + d = number of cases where one event has a characteristic and the other does not

$$S_m = \frac{a+b}{a+b+c+d}$$

problem: similarity due to missing items (such as two events have neither a gun nor fingerprints)

alternative: Jaccard's coefficient (eliminates the "negative matches")

$$S_m = \frac{a}{a+c+d}$$

Distance measures:
1) Euclidian -- uses square root

$$d_{ij} = \sqrt{\Sigma(x_i - x_j)^2}$$

2) City Block -- uses absolute value

$$d_{ij} = \Sigma\left|x_i - x_j\right|$$

3) Minkowski -- raised to r power, then 1/r

$$d_{ij} = \left(\Sigma\left|x_i - x_j\right|^r\right)^{1/r}$$

4) Mahalanobis -- distance squared; takes correlation into account

$$d_{ij}^{\ 2} = (x_i - x_j)' \Sigma^{-1} (x_i - x_j)$$

Problem with distance measures -- no inherent scale

Gower's Coefficient

d = dichotomous (y/n)
o = ordinal (more/less)
q = quantitative
w_o = # of ordinal characteristics used

$$G_s = \frac{\Sigma S_{dij} + \Sigma S_{oij} + \Sigma S_{qij}}{\Sigma w_j + \Sigma w_o}$$

Sara McLafferty, Hunter College

Spatial clustering -- What makes a hot spot hot?

- locations
- intensity of crime in hot spots
- statistical criterion for judging significance

How does one distinguish a hot spot from a random cluster?

- relationship of distances of people and property -- "underlying fabric"

Background

- Point-pattern analysis (Clark & Evans, 1950's ecology texts)
- Spatial statistics (Ripley, Diggle, Cresie)
- Spatial epidemiology (Clustering studies - Jacquez, Kulldorf)
- Geographical analysis (Openshaw, Bailey&Gatrell, Rushton (Geography)
- Crime Analysis (Blocks, LeBeau)

Methods

- Kernel smoothing
- Geographic Analysis Machine
- Besag and Newell

1) Kernel Smoothing -- on a map of points, create 3-D smoothed surface
With smoothing, one is now interested in *areas* rather than *points*
Move the Kernel circle (radius τ) across a window and estimate the intensity at various points. The following equation gives a measure of the intensity with a distance decay.

Estimated density at a point:

$$l(s) = \sum_{d_i < t} \frac{1}{t} k\left(\frac{d_i}{t}\right)$$

where:

$\lambda(s)$ = estimated density
d_i = distance from point i to point s
τ = bandwidth
k = kernel function (quartic is a typical kernel function)

Advantages of smoothed maps:
- visualization
- easy to compare across time periods
- avoid data agg.
- difference/ratio maps

Technical issues:
- choice of bandwidth (radius τ)
  - too small - spiky map, may replicate points
  - too large - flat map, no variation
- Use same bandwidth for entire map?
- Use smaller bandwidth in dense areas, larger in less dense areas?
- Boundary problem -- create a buffer zone to estimate, but not calculate, distances
- It is possible to use an adaptive kernel estimation where τ varies with λ.

Geographical Analysis Machine (GAM) -- Openshaw, et al
- generate grid points across study area
- generate circles of varying radii around points
- sum up # of cases and total population in each circle
- Poisson test -- Is # of cases significantly high?
- draw significant circles

Issues in GAM
- No statistical benchmark for # of significant circles
- False positives -- clusters that exist but are not real clusters in any meaningful sense
- Computational intensity

Besag & Newell method
An alternative to GAM - assumes that:
- Clusters will be found in areas of crime
- Point data for cases
- Areal data for population

Procedure:
For each case:
1. Find nearest areas that contain k cases
2. Compute total population of the areas
3. Poisson test for significance

The Besag & Newell method reduces the number of significance tests

Other methods:

Marshall (1991) - Journal of the Royal Statistical Society
Recent volumes of Statistics in Medicine

GIS:

Spatial Analyst for ArcView uses kernel smoothing
Vertical Mapper for MapInfo uses something very similar to kernel smoothing
Biomedware and S Plus Spatial Stats can be useful but must export to a GIS

Challenges

- Defining background distributions for study and comparison
- Significance tests are based on rare events and crimes are not always rare events
- Time dimension
- Combining spatial and a-spatial clustering methods
- Closer coupling of GIS and spatial analysis